

Identifying Built-up Areas for 2011 Census Outputs

Jenny Harding*, Bill South**, Mark Freeman*, Sheng Zhou*, Alex Babington*

* Ordnance Survey, GB

** Office for National Statistics, United Kingdom

Abstract. Datasets delimiting built-up areas, for use with England and Wales census data, had been produced for the census dates 1981 to 2001 by manual digitising. By the 2011 census, availability of attributed digital topographic polygon data presented opportunity for more efficient automated dataset creation. This paper discusses collaborative work between Ordnance Survey (the National Mapping Agency of Great Britain), and a government consortium led by the Office for National Statistics (ONS), to produce a fit for purpose and cost effective dataset of built-up areas for use with 2011 census data for England and Wales. Research challenges centred around whether an automated approach could be consistent with guidelines used to produce previous epochs of the dataset and meet the needs of diverse uses for built-up areas data.

Keywords: built-up area, census, context of use

1. Introduction

This paper discusses collaborative work between Ordnance Survey (the National Mapping Agency of Great Britain), and a government consortium led by the Office for National Statistics (ONS), to produce a fit for purpose and cost effective dataset of built-up areas for use with 2011 census data for England and Wales. Built-up areas data had previously been produced by manual digitising, but availability of attributed digital topographic polygon data presented opportunity for more efficient automated dataset creation. Research challenges centred around whether an automated approach could be consistent with guidelines used to produce previous epochs of the dataset (CLG 2001) and meet the needs of diverse uses for built-up areas data.

1.1. Background

The ONS has produced census data referenced to built-up areas in England and Wales every 10 years since 1981¹. This provides detailed information on settlements of all sizes, from villages to cities, and allows comparison between urban and non-urban populations at local authority level. Boundaries of the built-up areas need to be identified to enable these census outputs, besides being a pre-requisite for creation of the rural-urban classification for England and Wales. Uses of the data include statistical analysis and reporting, the development of policy, monitoring and planning by a wide range of users in public sector bodies, businesses and academia.

Ordnance Survey captured built-up area data for the 1981, 1991 and 2001 census by manually digitising extents from 1:10,000 scale mapping in accordance with guidelines provided by government stakeholders (CLG 2001) in which urban areas are defined as land which is 'irreversibly urban in character'. Included are areas of built-up land with a minimum area of 20 hectares (200,000 m²), while settlements separated by less than 200 metres are linked, and large conurbations are sub-divided.

Urban development since 2001 meant a revised dataset was needed for use with 2011 census data. A cross-government project team was set up in January 2011 to assess the feasibility of creating built-up areas data for England and Wales for 2011 by using an automated methodology and OS MasterMap[®] source data. Output data needed to be relevant and fit for purpose for users interested in understanding and analysing urban (and rural) issues.

1.2. Motivations for developing a new way to produce built-up areas data

Manual digitising was resource intensive and made the dataset expensive to produce. It was also naturally subject to some spatial inconsistency in capture between digitising operators. Advances in data structures and analytical tools since 2001, offered the possibility to automate all or part of the process, creating data with improved efficiency, consistency and transparency of method.

An important consideration was that 2011 statistical outputs should be compatible with previous data for built-up areas used with census data.

¹ For 2011 the dataset was renamed built-up areas, prior to that it had been known as urban areas

Therefore, the approach for delimiting built-up areas was to be based on capture guidelines used in 2001, though these were defined before automation was feasible and therefore not sufficiently specific for a complete algorithmic interpretation.

2. Approach

2.1. User needs survey

The census data for the built-up areas, the digital boundaries, and the rural-urban classification are used for a variety of purposes across local and central government organisations in England and Wales. A survey representing 9 main public sector uses for existing built-up areas data was carried out in order to provide wider user-focused context for evaluating technical options and addressing specification questions

Participants were recruited who had first hand knowledge of the dataset use within their organisation. A short telephone interview format was designed to explore context for principal tasks using the dataset, incorporating questions on key areas from 'context of use analysis' approaches originating in User Centred Design (see for example within ISO 9241-11). These covered: overall purpose of use and expertise involved; task success criteria; technical constraints; frequency of dataset use and motivations to use the dataset; high level specification needs such as for accuracy, currency and level of detail to support the task; coverage required; linkages to other data and priorities for dataset improvement.

Context of use results summary

The urban settlements dataset is used as an input to a wide range of tasks which could be broadly categorised as:

- Reporting statistics (socio-economic, population, environmental etc) by urban area
- Analysis and reporting of change over time in urban/rural areas and characteristics relating to them
- Spatial analyses and modelling relating to urban/rural areas
- Assessment/application of government funding in relation to urban/rural attributes
- Producing contextual maps (using the dataset as a backdrop)

People directly involved in these tasks are usually GIS and statistics experts and may involve other expertise, e.g. social/economic/demographic researchers, economists, planners, policy makers

The results of the survey highlighted some common areas that users felt were important in creation of the dataset. The three most frequently highlighted factors were:

- Spatial consistency (of built-up area definition)
- Reliability, credibility and robustness of method and trust in the dataset
- Temporal consistency of the dataset creation method for analyses over time (or at least clear explanation for any differences between dataset releases)

Built-up areas were most often required as polygon data for the types of analysis carried out. Many of the motivations for use of the dataset in the different task contexts stem from its usage with ONS census outputs. Other motivating aspects included the need to be consistent with other government users and to ensure that built-up areas are a geographic concept that most people can relate to.

In terms of data quality, the most critical accuracy considerations were for consistency of the data in relation to rules application in the method used; clarity on what the polygons represent and on limitations of the data and methodology. Increased update frequency of the dataset would benefit some users, though synchronicity with census data or other demographic data was important in many use contexts. Some users could also benefit from a degree of flexibility over the minimum built-up area size included in the dataset.

2.2. Technical Approaches investigated

Development of technical options included an automated algorithm identifying built-up area extents from polygon data for topographic features and an automated solution analysing land cover attribution and creating built-up area extents from grid based results. In both approaches the main base data input was that used for Ordnance Survey's largest scale vector topographic dataset, OS MasterMap® Topography Layer.

Automated algorithm based on topographic polygons

This approach used polygon geometry and attribution from the database for the large scales vector data as input to an algorithm to build built-up area

polygons from relevant underlying topographic feature polygons. It builds on research previously developed by Chaudhry and Mackaness (2008).

In summary, the prototyped algorithm first partitioned data from the source national database into processing chunks using motorways, roads with 'A road' classification and Mean High Water alignments. Within partitioned areas, buildings and glass house polygons were selected, buffered outwards and aggregated where their buffered extents intersected. Aggregated extents were buffered inwards by the same distance to approximate their true aggregated extent. Boundaries were then refined to a degree by applying rules in accordance with the type of topographic area included. Further processing connected polygons within 50m of each other, using a triangulation based approach.

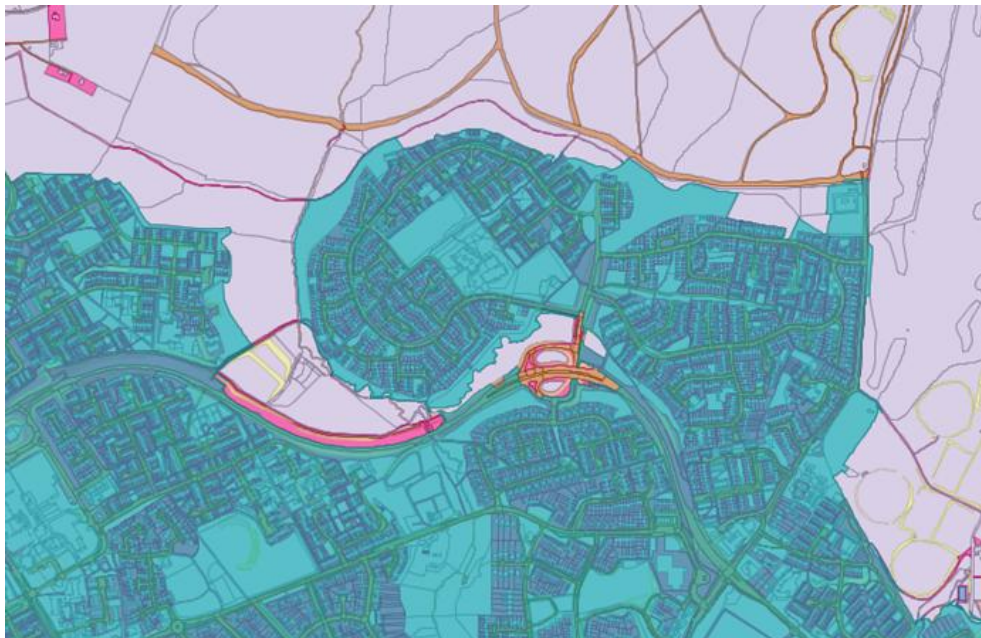


Figure 1. Polygon based delineation of built-up area (blue shaded area). Ordnance Survey © Crown Copyright. All rights reserved.

Resultant polygon boundaries output from the prototype algorithm followed topographic feature boundaries where these supported the rules base for defining the built-up area. In areas where there were sub 50m 'gaps' in built-up land cover the triangulation approach used by the algorithm provided a joining solution. Rules required to consistently manage the 50m adjacency requirements were difficult to define. Overall the approach resulted in a number of anomalous boundary alignments, for example in areas where features classed as 'built-up' had a large curtilage of vegetation. The pro-

cessing of very large areas of source polygon data (i.e. for regional or national coverage) using this approach is complex and requires a subsequent process to connect built-up areas, where appropriate, across the partitions created for processing purposes.

Automated algorithm based on values within grid squares

In this approach, the built-up area polygons were generated from an intermediate polygon dataset. This dataset was a tiling of square polygons, each attributed with a summary of their intersecting topographic features. The automated process of deriving built-up area polygons from the topographic features was as follows.

A tiling of square 50m x 50m polygons, aligned with the British National Grid system, was intersected with the OS MasterMap Topography Layer data. Within each grid square, the intersected features were interrogated for their land cover attributes and, where related to built-up area guidelines (CLG 2001), categorised into one of four nominal National Land Use Database (NLUD, ref. Harrison 2006) classes (i.e. those classifying: buildings and glass houses; metallised road surfaces; other areas of tarmac and concrete surfaces; residential gardens). Each polygon was then attributed with the percentage area covered by these NLUD classes; see example in Figure 2.



Figure 2. OS MasterMap Topography Layer extract overlaid by 50m grid, showing land cover attributes for one grid square (Ordnance Survey 2012). Ordnance Survey © Crown Copyright. All rights reserved.

Each grid polygon is categorised as 'built-up', or not, based on whether they intersect minimum threshold percentages of the land cover classes outlined above. One addition to this scheme is that grid polygons which only cross the threshold percentage for metallised surfaces are only included if they are adjacent to polygons which are classed as built-up based on one or more of the other NLUD percentages.

Grid cells adjacent to each other and which are categorised as built-up are then merged to form initial built-up area polygons. Enclosed holes in these areas (for example, parks that are fully surrounded by housing) are identified and merged into these polygons. Resultant polygons smaller than the desired minimum size (200 000 m²) and the 'holes' are discarded but could easily be re-captured in a supplementary dataset if there was user demand.

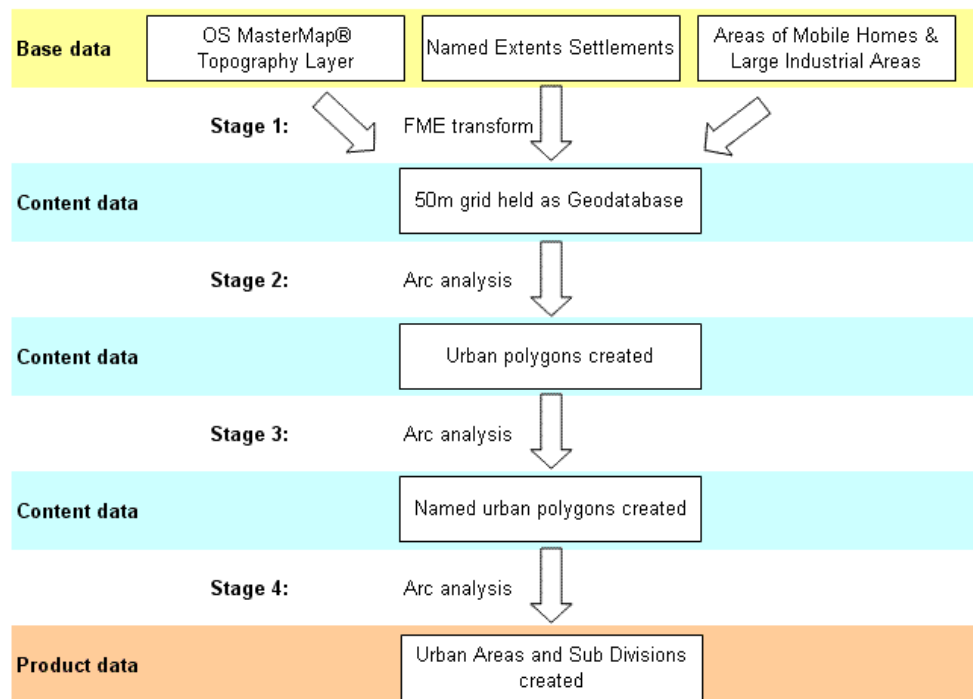


Figure 3. Process for creating grid based polygon data of built-up areas (Ordnance Survey 2012)

Resultant grid based polygons are inherently jagged in appearance but proved an efficient means of defining built-up area extents closely fitting the relevant underlying topographic detail, providing an attributed dataset for analytical uses. Fig. 4 shows an extract of grid based built-up area polygons which compares well with the built up area polygons for the same area as created by manual digitising in 2001 (Fig. 5). Differences in detail arise due

to consistent application of the rules base in the automated approach, as compared to manual interpretation of the original guidance.

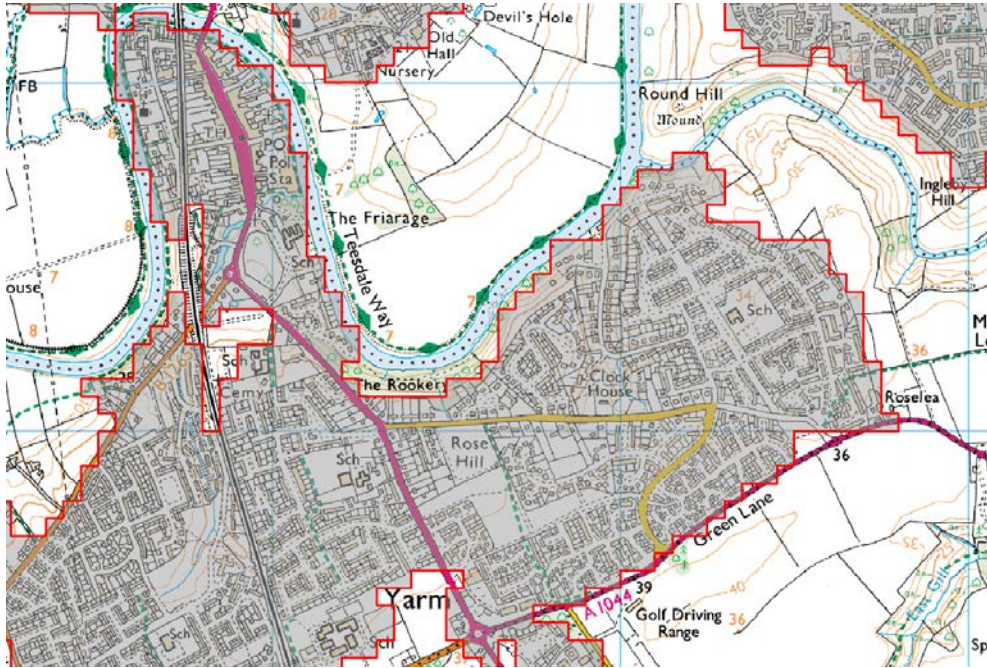
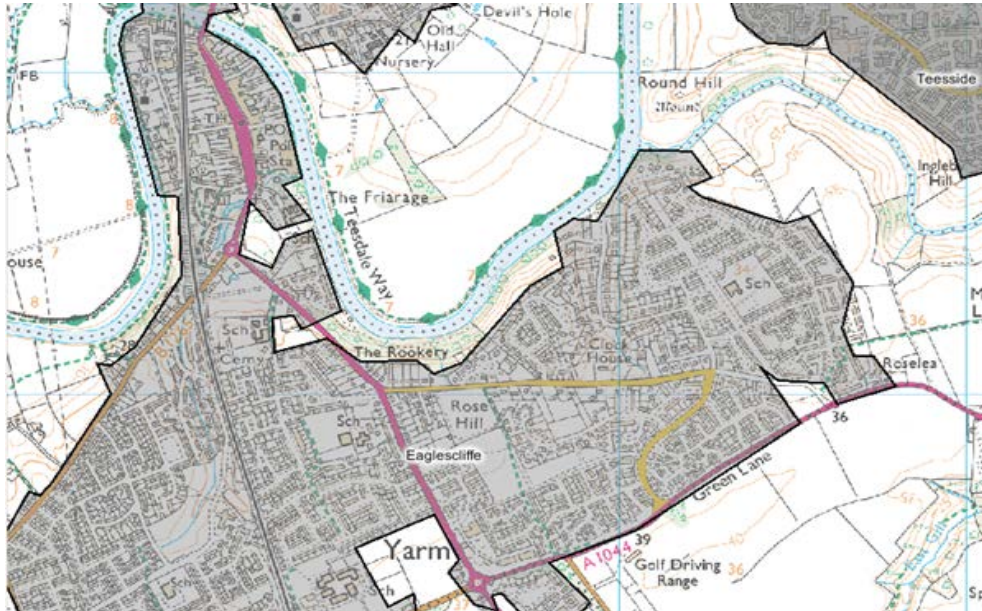


Figure 4. Extract from grid based polygon boundaries of the built-up area, overlying topographic detail from 1:25000 OS Explorer Map. Ordnance Survey © Crown Copyright. All rights reserved.



3. Discussion

3.1. Technical outcome and source data usability

Taking into account results of the user survey, relative costs of the technical options and preliminary analysis of statistical impact, the preferred approach was automation based on land cover attribution in 50m grid squares.

In addition to the automated process, a limited amount of manual capture from aerial imagery was required to complete the dataset. This was to enable the inclusion of large industrial sites, with typically mixed land cover types, and large areas of mobile homes. These are not sufficiently attributed in the source data for automatic classification as built-up land, but are required to be included in order to be consistent with previous versions of the dataset. A further rules based approach is used to link or subdivide polygons and apply names (for city, town etc.) to the resultant polygons from a separate names database. The output dataset is supplied as attributed vector polygons in Shapefile format. To achieve fully automated production of the dataset would require additional explicit attribution in the source data

to identify all land use types considered as 'built-up' in accordance with the guidelines (CLG 2001).

The actual built-up area polygons produced by the different approaches, vary in details of their geometry. The gridded polygons, while closely hugging the visual extent of built-up areas, do not trace real world features, while the polygon based approach traces alignment of topographic features at the perimeter of built-up areas but in places uses interpolation to approximate the perimeter where no topographic feature exists. To be practicable, the manual digitising approach used smaller scale source data (1:10000) and is subject to differences in interpretation by operators, both factors resulting in differences in built-up area geometry.

In comparison with previous productions of the dataset, the nature of the automated production methods will mean there are differences between the manually created 2001 dataset and an algorithmic based solution. However looking to future creations, automated approaches offer improved spatial and temporal consistency in the application of a rules base. For the key users of the dataset the degree of potential difference in data epochs due to change of methodology was felt to be acceptable.

3.2. Benefits of the grid based approach

The automated grid based approach proved effective in implementing the specification guidelines and addressing the critical factors highlighted in the user survey. It has the following benefits:

- Enhanced overall spatial consistency compared to manual production
- Improved credibility and transparency through clear metadata and documentation of the specification and method used
- Improved temporal consistency in subsequent generations of the dataset.
- Repeatability - the relative ease of re-running the automated process offers potential for more frequent updates of the dataset.
- Potential for additional data (for example, built-up areas smaller than the current 20 ha threshold and 'holes') to be created at limited additional cost

A negative aspect of the grid based polygons is their jagged appearance (50m grid resolution) if used for graphic representation purposes, in particular at scales larger than 1:50 000. This may be an issue for some users from an aesthetic perspective.

3.3. Census statistics and grid based built-up areas data

2011 Output Areas (OAs) (ONS 2010) have been best-fit to the built-up area and built-up areas sub-division boundaries by ONS using a population-weighted centroid methodology. Approximately 5,500 built-up areas and 1,700 built-up area sub-divisions were found to contain a population in this process (i.e. were assigned one or more OA). This will allow a range of 2011 Census statistics to be produced for the areas:

- i. Counts for built-up areas (with sub-divisions) for all areas
- ii. Counts for built-up areas (including those classed as 'urban' in the rural-urban definition) and remainder areas by local authority
- iii. Summary tables, population based, ranked as follows:
 - 1,000,000 and above
 - 500,000 - 999,999
 - 200,000 - 499,999
 - 100,000 - 199,999
 - 50,000 - 99,999
 - 20,000 - 49,999
 - 10,000 - 19,999
 - 5,000 - 9,999
 - 2,000 - 4,999
 - 1,500 - 1,999

Some built-up areas/sub-divisions did not contain population (i.e. were not assigned to an OA), for example large industrial sites. These will not appear in the Census outputs but will be retained in the digital boundaries dataset.

4. Conclusion

Through working closely with users in the project group on iterative development of the dataset, a cost effective and fit for purpose solution was achieved. With the grid based process, an urban areas dataset could be produced for England and Wales in days rather than with months of manual digitising effort. It offers the additional benefits of spatial and temporal consistency in the automated application of its rules base, being easily repeatable and giving potential for more frequent updates. Based on OS MasterMap large scales data, resolution is improved compared to previous versions based on 1:10,000 mapping, and the algorithms have flexibility to include built-up areas smaller than the current threshold of 20 hectares. Each of these factors, together with the transparency of the data creation approach, meets key criteria identified in the user survey.

Acknowledgements

The authors gratefully acknowledge very significant contributions to this work of others in the government consortium for the project: Alistair Edwardes (formerly of DCLG), Carol Hryniewicz (DEFRA), Justin Martin (DEFRA), Stuart Neil (WG); survey participants; and in the OS team: Pete Booth, Owain Hale-Heighway, Phil Wyndham and especially Rob Gower.

References

Chaudhry O, Mackaness W (2008) Automatic Identification of Urban Settlement Boundaries for Multiple Representation Databases. *Computers, Environment and Urban Systems* 32: 95-109.

CLG (2001) Urban Settlements 2001 Data Methodology Guide. Dept. For Communities and Local Government

Harrison, A. (2006) National Land Use Database: Land Use and Land Cover Classification, version 4.4. Office of the Deputy Prime Minister. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/11493/144275.pdf Accessed 26 March 2013

ISO (1998) ISO 9241-11: 1998 Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) - Part 11: Guidance on Usability.

ONS (2010), Geography Policy for National Statistics, <http://www.ons.gov.uk/ons/guide-method/geography/geographic-policy/index.html>. Accessed 26 March 2013

Ordnance Survey (2012) Built-up Areas Core Specification. v2.0. Ordnance Survey